Stopping Stochastic Approximation

David W. Hutchison
Mathematical Sciences
The Johns Hopkins University
3400 North Charles Street
Baltimore, MD 21218

James C. Spall
Applied Physics Laboratory
The Johns Hopkins University
11100 Johns Hopkins Road
Laurel, MD 20723-6099

ABSTRACT

The practical application of stochastic approximation methods require a reliable means to stop the iterative process when the estimate is close to the optimal value or when further improvement of the estimate is doubtful. Conventional ideas on stopping stochastic algorithms employ probabilistic criteria based on the asymptotic distribution of the stochastic approximation process, often with the parameters of the distribution determined by sequential estimation. Difficulties may arise when this approach is applied to small (finite) samples. We propose a different approach that uses the notion of an idealized process as a companion to the stochastic approximation. A discussion of this approach to stopping stochastic approximation is offered in the context of a simple example, including some empirical results.

KEYWORDS: *adaptive and learning control systems, stochastic approximation, stopping*

1. INTRODUCTION

Introduced by Robbins and Monro [11] in 1951, stochastic approximation is the adaptation of iterative optimization and root-finding methods to stochastic problems. Robbins and Monro proved convergence in probability of a sequence of estimates to an optimal value, results which have since been strengthened to almost sure convergence. Yet these results only hold asymptotically. In practical applications the process must have a stopping condition, and some statements on when to stop the procedure or about the quality of the solution obtained must be made.

The need for a stopping rule for stochastic approximation was recognized by Kiefer and Wolfowitz [4]. Since that time this issue has been extensively studied. Burkholder [1] proposed estimators for the asymptotic distribution and derived sufficient conditions for those estimators to converge almost surely. Chow and Robbins [2] developed a method to sequentially determine a bound on the mean of a continuous random variable with unknown variance. Recognizing the applicability of this procedure to the problem of stopping stochastic approximation they suggested the rule: stop as soon as the length of the confidence interval based on asymptotic normality of the sample means is smaller than 2δ for some $\delta > 0$, or, equivalently, stop as soon as the estimated standard deviation of the sample mean is sufficiently small.

Since this initial work much of the effort in stopping stochastic approximation has been on the estimation of the parameters of the asymptotic distribution in order to apply one of the above criteria. Sielken [12] and Stroup and Braun [16] both improved on the pioneering work of Burkholder, applying sequential estimation to calculate estimators and develop confidence intervals in one dimension. These results were later extended to the multi-dimensional case [9, 17]. The conditions that guarantee the asymptotic validity for sequentially estimated parameters were established in [3].

2. PROBLEM FORMULATION

2.1 *The Stochastic Approximation Process*

We consider only the unconstrained case. Suppose $\theta \in \mathbb{R}^p$ is a vector with components representing control parameters. Let $Q(\theta, \omega)$ denote the observed response as a function of θ and the stochastic effect ω. The function of interest is $L(\theta) = E[Q(\theta, \omega)]$, the expected response at θ. The objective is to find the value θ^* that minimizes $L(\theta)$:

$$\theta^* = \arg\min_{\theta} L(\theta).$$

If the parameters θ are continuous and the solution space can be assumed closed and convex, then the problem lends itself to solution with a gradient-based optimization method. We choose an initial estimate $\hat{\theta}_0$ and update it with the following scheme:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k G_k(\hat{\theta}_k, \omega) \tag{1}$$

where $G_k(\hat{\theta}_k, \omega) \in \mathbb{R}^p$ is some (noisy) input information related to the gradient of the process being studied [4, 11]. A general discussion of the stochastic approximation method may be found in [15]. The asymptotic properties of $\hat{\theta}_k$ are well-known [6], and under relatively mild conditions the sequence of iterates generated by (1) converges to θ^* almost surely. See, for example, [8] or [5]. Additionally, when properly scaled, the distribution function of the iterates, which we denote F_k, is asymptotically normal. We denote the asymptotic distribution by F^*.

2.2 *Stopping the Approximation Process*

Since the estimates $\hat{\theta}_k$ are random, the preferred approach is to consider the probability that some tolerance conditions are met. Ideally we prefer conditions of the form

$$P_{\hat{\theta}_k}\left(\| \hat{\theta}_k - \theta^* \| < \delta \;\middle|\; \hat{\theta}_0\right) \geq 1 - \alpha \tag{2a}$$

$$P_{\hat{\theta}_k}\left(|L(\hat{\theta}_k) - L(\theta^*)| \; < \delta \;\middle|\; \hat{\theta}_0\right) \geq 1 - \alpha \tag{2b}$$

(see also [9, 10]). In a practical sense these conditions would allow us to compute the $1 - \alpha$ confidence ellipse about $\hat{\theta}_k$. Given an $\alpha \in [0, 1]$ and $\delta > 0$, we stop at time $\kappa(\alpha, \delta)$ equal to the smallest k such that either condition in (2) is true. To formalize the problem, a customary approach is to define $B(\hat{\theta}_k, \delta)$, a ball of radius δ about $\hat{\theta}_k$, and let κ be the first time the confidence ellipse based on (2) is contained within the ball.

Unfortunately, direct calculation of the probabilities in (2) is not possible since the distribution of $\hat{\theta}_k$ (even the distributional form) is generally not known. Some manner of estimation of the distribution function F_k is essential, and if convergence in distribution is sufficiently fast, one solution is to use F^* in lieu of F_k and estimate the parameters of F^*. Since the form of the distribution F^* is known, this is a well-defined problem.

The method requires knowledge of the covariance matrix Σ of the asymptotic normal distribution, and the Hessian at the optimal point, $H(\theta^*)$, of the underlying function. These matrices are not commonly available, so the usual procedure is to estimate them sequentially as part of the iteration. Given initial estimates for Σ and $H(\theta^*)$, these estimates are then updated at each step (or perhaps every m steps) as the iterative process proceeds, and the estimators are asymptotically correct.

This approach is less satisfactory for finite-sample (non-asymptotic) stochastic approximation. The difficulties extend beyond the fact that there may not be sufficiently many iterations to compute reliable estimates for Σ and $H(\theta^*)$. Finite-sample behavior could differ significantly from asymptotic behavior. This is a difficulty with the basic assumption that F^* can replace F_k in the calculation of the confidence ellipses. The assumption is a good one only asymptotically.

3. IDEALIZED PROCESSES

The concept of idealized processes is to develop a parameterized companion to the original process whose properties are known when the parameter is some positive number (giving the idealized process), and whose behavior reflects that of the original when the parameter is zero. The expectation is that conclusions we draw about the idealized process can be related to the original process in some way determined by the parameter.

This is a relatively new idea, and the theoretical justification for such a procedure is incomplete. The applicability of this method, however, has been shown for parameter estimation in maximum likelihood estimation problems, among others [13, 14]. We intend to demonstrate by example that this method can be used to stop stochastic approximation processes as well.

The idea of using idealized processes for parameter estimation was advanced by Spall [13, 14]. Spall's formulation sought an estimate $\hat{\theta}$ for a parameter vector θ from a set of data whose distribution depended on θ and a known scalar ϵ. When the sample is small, it is difficult to say much about the probabilities of $\hat{\theta}$ because the distributions are unknown. One approach is to construct a parameterized process producing statistically similar data and resulting in an estimate $\tilde{\theta}$ where the probabilities of $\tilde{\theta}$ are known, and then look for conditions where the probabilities of $\hat{\theta}$ are close to those of $\tilde{\theta}$ irrespective of the sample size.

We apply the same principle, but to the sequence of iterates from a stochastic approximation process. The idea is to establish conditions under which the probabilities of $\hat{\theta}_k$ are close to those of $\tilde{\theta}_k$, which are known. The resulting information is used to decide whether to stop the process or, if stopped, to determine the probability of being close to θ^*.

The stochastic approximation process in its most general form is

$$\hat{\theta}_{k+1} = T_k(\hat{\theta}_k, \omega) \tag{3a}$$

where T_k is some transformation process and ω denotes the random component which manifests itself as noise in the measurements of the loss function or its gradient. In the case of equation (1), T_k is a nonlinear operator with $T_k(\hat{\theta}_k, \omega) = \hat{\theta}_k - \alpha_k G_k(\hat{\theta}_k, \omega)$, where $G_k(\hat{\theta}_k, \omega)$ is a noisy estimate of the gradient of $L(\hat{\theta}_k)$. We parameterize the transformation with a scalar η:

$$\hat{\theta}_{k+1} = T_k(\hat{\theta}_k, \omega; \eta). \tag{3b}$$

For some $\eta = \eta_0 > 0$ the sequence of iterates produced by (3b) is the same[1] as (3a). For $\eta = 0$ (3b) produces a sequence of idealized iterates in the sense that the probabilities for the iterates are known for each k. For convenience we denote the sequence of iterates from the idealized process by $\tilde{\theta}_k$, and the distribution function of these iterates by G_k. (Note: when there is no confusion, we will often drop the ω from the expression for T_k.)

The idealized process $T_k(\theta; 0)$ could represent a simplified process that converges to the same θ^*, but more frequently the idealized process merely mimics some of the asymptotic properties of the true process. It cannot generally be shown (nor is it necessary to show) that $\tilde{\theta}_k \to \theta^*$. It is only necessary that the parameterized transformation process generate a sequence of dependent observations with distributional properties (other than location) that are similar to those of $\hat{\theta}_k$.

In general, it is not easy to determine a suitable parameterization for a general process. In the example that follows we assume the parameterization is known for analytical purposes, but further work in necessary before a systematic method of parameterization can be presented.

The justification for this approach is found in the use of the Skorokhod representation theorem to map the original process into another process on a different space where an analysis of the properties of the process is easier.

The general approach is to simplify the transformation process to generate an idealized process that is nearly identical up to some order. If the differences are small enough to be ignored, then the probability distributions should be close as well.

Formalizing this argument gives the following theorem: Let $\hat{\theta}_k = T_k(\hat{\theta}_{k-1}; \eta_0)$ be a stochastic approximation process with mean sequence $\bar{\theta}_k$ and covariance process Σ_k. Let $\tilde{\theta}_k = T_k(\tilde{\theta}_{k-1}; 0)$ be a linear idealized process relative to $\hat{\theta}_k$ (linearized about θ_0), and let $\tilde{\theta}_k$ have covariance process $\tilde{\Sigma}_k$. Assume the conditions required for the convergence of $\hat{\theta}_k \to \theta^*$ hold. Let $S(\theta)$ be any symmetric region about the point θ. Let $F_k(\bar{\theta}_k, \Sigma_k)$ be the true distribution of $\hat{\theta}_k$, and let $G_k(\bar{\theta}_k, \tilde{\Sigma}_k)$ be the idealized distribution of $\hat{\theta}_k$. Then

$$P_{F_k}\left(\hat{\theta}_k \in S(\theta^*)\right) - P_{G_k}\left(\hat{\theta}_k \in S(\theta^*)\right) = O\left\| \hat{\theta}_k - \theta_0 \right\|^2. \tag{4}$$

4. EXAMPLE

We take as an example the idealized process formed by linearizing the gradient. This results in an autoregressive process whose properties can be determined analytically. Consider the process given by (1) where $G_k(\theta, \omega) = g(\theta) + e_k(\omega)$ is the noisy gradient of $L(\theta)$. Here $T_k(\theta, \omega) = \theta - a_k g(\theta) - a_k e_k(\omega)$. The first order Taylor expansion of the gradient about a point θ_0 is

$$g(\theta) = g(\theta_0) + H(\theta_0)(\theta - \theta_0) + O(\| \theta - \theta_0\|^2 I_p).$$

The notation $H(\theta_0) = \nabla g(\theta_0)$ is shorthand for $\nabla g(\theta)|_{\theta=\theta_0}$, the Jacobian of g evaluated at θ_0. Since g is the gradient of L, the Jacobian of g is the Hessian of L, and we use the function H to denote this. The natural parameterization is

$$\begin{aligned} T_k(\theta; \eta) = \theta - a_k g(\theta_0) - a_k H(\theta_0)(\theta - \theta_0) \\ + \eta\, O(\|\theta - \theta_0\|^2 I_p) - a_k e_k. \end{aligned} \tag{5}$$

When $\eta = \eta_0$ (for some η_0) we have the true process (1) with $G_k(\theta, \omega) = g(\theta) + e_k(\omega)$. When $\eta = 0$ we have the approximation process

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - a_k g(\theta_0) - a_k H(\theta_0)(\tilde{\theta}_k - \theta_0) - a_k e_k. \tag{6}$$

Since $T_k(\theta_k; 0)$ is linear in the random components, it follows from repeated substitution that each iterate $\tilde{\theta}_{k+1}$ in 6 may be expressed as a deterministic variable (depending on k) plus the sum of multiples of the random variables $\{e_0, e_1, \ldots, e_k\}$. If we can assume nice behavior of the e_k, then the distribution function $F_{\tilde{\theta}_k}$ is known for any k. This distribution is used to compute $\kappa(\alpha, \delta)$ for some α and δ according to (2).

We illustrate this idea by using a function from the Moré et al. [7] suite of optimization problems, the so-called variably dimensioned function in two dimensions, to generate stochastic data and to identify the linear idealized process as a test of the procedure. Let $\theta = [t_1\ t_2]^{\mathrm{T}} \in \mathbb{R}^2$ and $L_{VD} : \mathbb{R}^2 \to \mathbb{R}$. Then the variably dimensioned function is defined as

$$\begin{aligned} L_{VD}(\theta) = (t_1 - 1)^2 + (t_2 - 1)^2 \\ + (t_1 + 2t_2 - 3)^2(1 + (t_1 + 2t_2 - 3)^2). \end{aligned}$$

The gradient is

$$g_{VD}(\theta) = \begin{bmatrix} 4t_1 + 4t_2 + 4(t_1 + 2t_2 - 3)^3 - 8 \\ 4t_1 + 10t_2 + 8(t_1 + 2t_2 - 3)^3 - 14 \end{bmatrix}$$

[1] By the same we mean statistically indistinguishable.

This function satisfies the conditions for convergence of (1), and there is a unique global minimum located at $\theta^* = [1\ 1]^{\mathrm{T}}$.

Suppose the form of the loss function L_{VD} is not known, but we are able to provide inputs θ and observe the noisy gradient. We assume the components of the noise e_k are independent and normally distributed with mean zero and variance ${\sigma_k}^2$. Then we model the process as follows: for a sequence of inputs $\{\theta_k\}$ we have a sequence of observations $\{Y_k\}$ generated by $Y_k(\theta_k, \omega) = g_{VD}(\theta_k) + e_k(\omega)$. Let $\hat{\theta}_k = [\hat{t}_1\ \hat{t}_2]^{\mathrm{T}}$ and $\tilde{\theta}_k = [\tilde{t}_1\ \tilde{t}_2]^{\mathrm{T}}$. Using Robbins-Monro iteration the true ($\eta = \eta_0$) process is

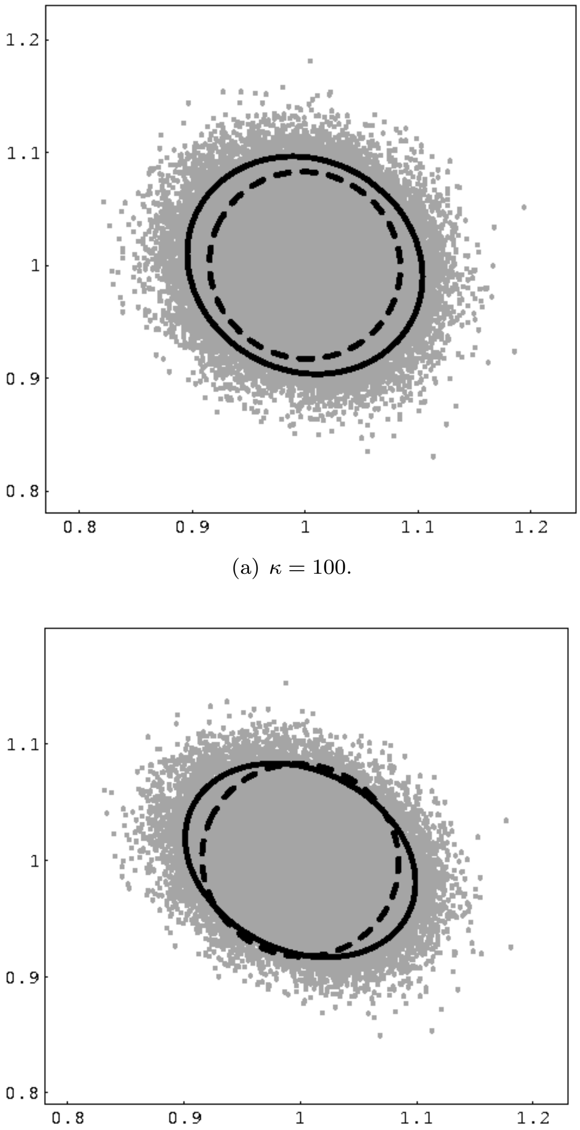
$$\begin{aligned}\hat{\theta}_{k+1} = \quad & \hat{\theta}_k - a_k g_{VD}(\hat{\theta}_k) - a_k e_k \\ = \quad & \begin{bmatrix} \hat{t}_1 \\ \hat{t}_2 \end{bmatrix} \\ & -a_k \begin{bmatrix} 4\hat{t}_1 + 4\hat{t}_2 + 4(\hat{t}_1 + 2\hat{t}_2 - 3)^3 - 8 \\ 4\hat{t}_1 + 10\hat{t}_2 + 8(\hat{t}_1 + 2\hat{t}_2 - 3)^3 - 14 \end{bmatrix} \\ & -a_k e_k \end{aligned}$$

and the idealized process is

$$\tilde{\theta}_{k+1} = \begin{bmatrix} \tilde{t}_1 \\ \tilde{t}_2 \end{bmatrix} - a_k \begin{bmatrix} 4\tilde{t}_1 + 4\tilde{t}_2 - 8 \\ 4\tilde{t}_1 + 10\tilde{t}_2 - 14 \end{bmatrix} - a_k e_k.$$

The distribution function F_k is unknown, and impossible to calculate. However, it can be approximated using a Monte Carlo experiment. For this example, the distribution function for the idealized process, $\tilde{F}_k$, is the sum of bivariate normal random variables, and can be calculated exactly. The Robbins-Monro algorithm with step size $a_k = a/k$ was stopped after κ steps using $\hat{\theta}_0 = \theta^*$ and $\sigma_k = 10$. The scatterplots in Figure 1 were generated by taking 100,000 such runs.

It is apparent from the plots that G_k is a better approximation to F_k than F^*, even for high iteration counts. One measure of how well G_k approximates F_k when $\hat{\theta}_0 = \theta^*$ is to look at the Kullback-Leibler distance between the sample probability mass function and values of the distribution of the idealized process. In this example, computations show that the Kullback-Leibler distance is actually decreasing with increasing iterations. This is mostly an artifact of choosing to expand about θ^*. An expansion about the initial point $\hat{\theta}_0$, say, leads to an approximation that is initially good, but gets worse as the iteration progresses. In this instance one might ”restart” the process by relinearizing the gradient (at the current point, say) and continuing the process from there.



(a) $\kappa = 100$.

(b) $\kappa = 10,000$.

Figure 1: The figures above show a scatterplot of $\hat{\theta}_\kappa$ overlayed with the approximate 95th percentile ellipse of the distribution (solid ellipse) for the value of κ shown when $\hat{\theta}_0 = \theta^*$. The 95th percentile ellipse for $\tilde{\theta}_\kappa$ is indistinguishable from the solid ellipse in this scale, and therefore is not shown. The dashed ellipse represents the 95th percentile of the asymptotic distribution of $\hat{\theta}_\kappa$.

5. CONCLUSIONS

There currently is no systematic way to select the idealized process or to evaluate alternatives. The problem with the linearized transformation is that information on the Hessian is needed, and in a practical application it must be estimated from the observations. This was a drawback of the sequential estimation approach in small finite samples, and with a linearized transformation we cannot escape it here (though it may be better to use this information to estimate F_k rather than F^*). Other autoregressive processes show potential.

Additional advantages to this approach include relative computational efficiency compared to other methods and the potential to take into account in an explicit manner the impact of a poorly selected $\hat{\theta}_0$. The questions that remain point out the need for a more general theory for idealized processes.

References

[1] D. Burkholder. On a class of stochastic approximation procedures. *Annals of Mathematical Statistics*, 27:1044–1059, 1956.

[2] Y. Chow and H. Robbins. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Annals of Mathematical Statistics*, 36:457–462, 1965.

[3] P. W. Glynn and W. Whitt. The asymptotic validity of sequential stopping rules for stochastic simulations. *The Annals of Applied Probability*, 2(1):180–198, 1992.

[4] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[5] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York, 1997.

[6] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, NJ, 2 edition, 1999.

[7] J. J. Moré, B. S. Garbow, and K. E. Hillström. Testing unconstrained optimization software. *ACM Transactions on Mathematical Sciences*, 7(1):17–41, 1981.

[8] M. B. Nevel'son and R. Z. Has'minski. *Stochastic Approximation and Recursive Estimation*, volume 47 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1973.

[9] G. C. Pflug. Stepsize rules, stopping times, and their implementation in stochastic quasigradient algorithms. In Y. Ermoliev and R. J.-B. Wets, editors, *Numerical Techniques for Stochastic Optimization*, pages 353–372. Springer-Verlag, New York, 1988.

[10] G. C. Pflug. *Optimization of Stochastic Models*. Kluwer Academic Publishers, Boston, 1996.

[11] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[12] R. L. Sielken, Jr. Some stopping times for stochastic approximation procedures. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 27:79–86, 1973.

[13] J. C. Spall. Uncertainty bounds for parameter identification with small sample sizes. In *Proceedings of the 1995 IEEE Conference on Decision and Control*, pages 3504–3515, New Orleans, 1995. IEEE.

[14] J. C. Spall. Uncertainty bounds in parameter estimation with limited data. In M. Dror, P. L'Ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pages 685–710. Kluwer Academic Publishers, Boston, 2002.

[15] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, New York, 2003.

[16] D. F. Stroup and M. I. Braun. A new stopping rule for stochastic approximation. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 60:535–554, 1982.

[17] G. Yin. A stopping rule for the Robbins-Monro method. *Journal of Optimization Theory and Applications*, 67(1):151–173, 1990.